DLEP: A Deep Learning Model for Earthquake Prediction

Rui Li, Xiaobo Lu, Shuowei Li, Haipeng Yang, Jianfeng Qiu and Lei Zhang
Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230039, China
Email: ruili200009@126.com, e11714034@stu.ahu.edu.cn, shuowei20@outlook.com, haipengyang@126.com, qiujianf@ahu.edu.cn, zl@ahu.edu.cn (*corresponding author*)

***Abstract*—Earthquakes are one of the most costly natural disasters facing human beings, which happens without an explicit warning, therefore earthquake prediction becomes a very important and challenging task for humanity. Although many existing methods attempt to address this task, most of them use either seismic indicators (explicit features) designed by geologists, or feature vectors (implicit features) extracted by deep learning methods, to characterize an earthquake for earthquake prediction. The problem of combining these two kind of features to improve final earthquake prediction performance remains pretty much open. To this end, we propose a deep learning model named DLEP to effectively fuse the explicit and implicit features for accurate earthquake prediction. In DLEP, we adopt eight precursory pattern-based indicators as the explicit features, and use a convolutional neural network (CNN) to extract implicit features. Then, an attention-based strategy is suggested to fuse these two kinds of features well. In addition, a dynamic loss function is designed to deal with the category imbalance of seismic data. Finally, experimental results on eight datasets from different regions demonstrate the effectiveness of the proposed DLEP for earthquake prediction comparing to several state-of-the-art baselines.**

***Index Terms*—Earthquake prediction; Feature extraction; Explicit feature; Implicit feature; Deep learning method.**

I. Introduction

Earthquakes are one of the most devastating natural disasters in the world, which occur without an explicit warning and may cause serious injuries or loss of human lives. One of effective solutions for reducing earthquakes loss is the earthquake prediction, which aims to use the known earthquake data to specify three elements, namely when, where and the magnitude of the future earthquake. Therefore, effective earthquake prediction can reduce the earthquake damage to a large extent, which is of great significance to the country and society, and there has been an increasing interest and academic research on predicting seismic events.

In the effort to predict earthquakes such as the magnitude, time and location of the earthquakes, many researchers attempted to use physical methods to explain and describe earthquakes, and studied earthquake precursors through the study of geology [1]–[3]. In these works, researchers designed many indicators of earthquake, such as earthquake magnitude and intensity energy, as the explicit features of seismic events [4], [5]. The values of these indicators are calculated based on the data of site investigation, and different feature extraction methods based on precursory patterns were also proposed [6], [7], to predict the magnitude of earthquakes in the next period of time. For example, Zhang et al. [7] recently proposed a precursory pattern-based feature extraction method to enhance the performance of earthquake prediction, based on which the eight mathematical statistic features can be generated as seismic indicators. The experimental results on two historical earthquake records demonstrated the effectiveness of their precursory pattern-based features with the selected CART algorithm for earthquake prediction. In summary, explicit features extracted by humans have strong interpretability from the theoretical system, and can realistically describe an earthquake to some context. However, these manually designed features may fail to fully utilize information contained in seismic sequences.

To this end, some researchers try to establish the neural network model to predict earthquake without explicitly modeling features. For example, DeVries et al. [8] proposed a deep-learning approach to identify a static-stress-based criterion, which predicted aftershock locations without prior assumptions about fault orientation. In addition, Huang et al. [9] used the convolutional neural networks (CNN) to extract the implicit features from the geographic images marked with seismic information for large earthquake magnitude prediction in Taiwan. The experimental results demonstrated that implicit features can be used to find feasible solutions for earthquake prediction problems from another perspective. Although implicit features extracted by deep learning methods can fully utilize information contained in seismic sequences, they have weak interpretability from the theoretical system.

In summary, most of the existing works characterize the earthquake by using either seismic indicators (i.e. explicit features) designed by geologists or experts, or using feature vectors (i.e. implicit features) extracted by deep learning methods. In reality, for more accurate earthquake prediction, it is necessary and challenging to design novel model that can combine the advantages of explicit features and implicit features. Moreover, it is found that there are often serious category imbalance problem in seismic data, that is, the earthquakes with relatively high magnitude usually occupy a small part of the vast majority of seismic data while most of the earthquake events in the dataset are small-scale.

To solve the above challenges, in this paper, we propose a novel deep learning model for accurate earthquake prediction, which can effectively combine the explicit and implicit fea-

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

tures as well as deal with the category imbalance problem. To be specific, the seismic sequence is firstly divided into many representative learning samples and precursory patterns. Based on these patterns and samples, the eight mathematical statistics based earthquake indicators [7] are adopted as explicit features, while the implicit features are extracted by CNN with precursory patterns as input. Then, an attention-based strategy is suggested to combine the advantages of both explicit features and implicit features well. In addition, due to the small batch gradient descent method is adopted in the model optimization, we design a dynamic loss function to take both the population distribution of the samples and the distribution of each batch into account, to accommodate different training data, thus can solve the challenge of category imbalance in seismic data.

In summary, the contributions of this paper can be summarized as follows:

- We argue that the feature extraction methods used in previous earthquake prediction methods obtain explicit features by geologists and implicit features by deep learning methods individually, and lack a general model that can combine the advantages of both explicit features and implicit features.
- We propose a novel deep learning model named DLEP for earthquake prediction. In DLEP, the explicit features and implicit features are combined effectively by a suggested attention-based strategy. Furthermore, a dynamic loss function is also designed for dealing with the category imbalance problem of seismic data.
- We evaluate the effectiveness of our model DLEP comparing to state-of-the-art baselines, and the experimental results on eight datasets with different characteristics demonstrate the promising performance of the proposed DLEP, which indicates that the idea of fusing both explicit features and implicit features is an effective solution for accurate earthquake prediction.

The rest of this paper is organized as follows. Section II presents the preliminaries and related work. Section III describes the proposed model DLEP in detail. In Section IV, experimental results are presented and discussed, followed by a conclusion and future work in Section V.

II. The Preliminaries and Related Work

In this section, some preliminaries about problem definition and the structure of CNN are firstly described, and some related work about earthquake prediction is then introduced in detail.

A. Problem Formulation

Seismic data from around the world are presented as sequences of earthquake events, which contains much seismic information, such as latitude, longitude, magnitude and so on. In order to obtain the required samples of the training model, the work in [7] divides the raw seismic data into a set of fixed day time period N, and then defined the precursory pattern in each fixed time period. This kind of sequence segmentation method is simple and easy to understand, however, the result of

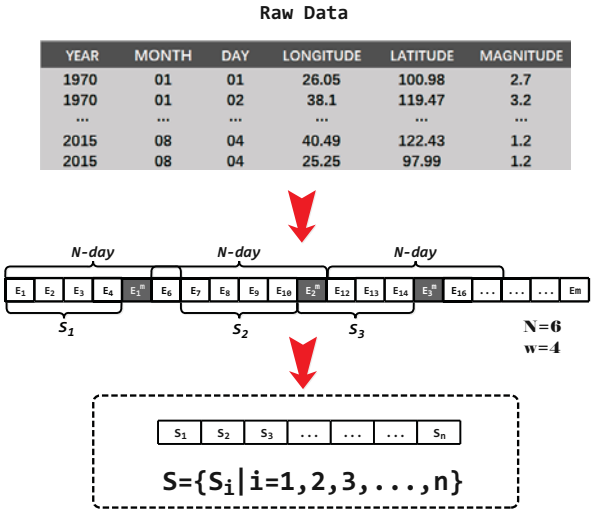

Fig. 1: The definition method of precursory pattern

division depends on the value of time period to a large extent. In other words, if the value is taken improperly, it is likely to miss some important information in the seismic sequence. To this end, we propose a new partitioning method for the definition of precursory patterns. To be specific, suppose the historical seismic record is denoted as θ, which is defined as:

$$\theta = \{E_i | i = 1, 2, \ldots, m\}$$

$$E_i = \langle a_{i1}, a_{i2}, \ldots, a_{ij}, \ldots, a_{it} \rangle$$

where E_i is the i-th earthquake event and the sequence θ includes m earthquake events. Besides, a_{ij} is the j-th attribute of the i-th earthquake event and t is the total number of attributes for describing an event. In real earthquake data, each event E_i consists of at least two seismic attributes, such as earthquake magnitude and occur time. Given a fixed pre-defined N days, the earthquake event with the largest magnitude that occurred in each N days period is called main shock. The main shock is denoted as E_i^m $(1 \leq i \leq n)$, where n is the total number of main shocks in θ. In addition, the w events before each main shock E_i^m are formed as one precursory pattern P_i. Formally, the raw sequence θ can be segmented as a sample set S:

$$S = \{S_i | i = 1, 2, \ldots, n\}$$

$$S_i = \langle P_i, E_i^m \rangle$$

$$P_i = \langle E_{i-1}^m, E_{i-2}^m, \ldots, E_{i-w}^m | i > w \rangle$$

where each sample S_i consists of the main shock E_i^m and its precursory pattern P_i.

For example, as shown in Fig. 1, suppose N = 6 and w = 4, the first N earthquake events is selected in the raw data sequence, and the event E_1^m with the largest magnitude is the main shock in the first N-events. Then, the w events before E_1^m make up the precursory pattern $P_1 = \{E_1, E_2, E_3, E_4\}$. Thus, P_1 and E_1^m make up the first sample S_1. Then, the event subsequence with length N after the E_1^m is considered as the

second N-event and this period is start from E_6. We can find that E_2^m is the main shock in the second N-events. Similarly, the w events before E_2^m make up the precursory pattern $P_2 = \{E_7, E_8, E_9, E_{10}\}$. Thus, P_2 and E_2^m make up the second sample S_2. The remaining samples $S_3, ..., S_n$ are obtained in the similar way. It can be found that our partitioning method is based on the main shocks and dynamically segments the raw data sequence to extract precursory patterns and the samples. The partition result does not depend too much on the value of N, and strictly keep the events after each main shock makes higher utilization rate of the raw data.

Based on the above notations, the earthquake prediction problem is formally defined as:

Definition 1: (***Earthquake Prediction***) Given the historical earthquake sequence θ, the pre-defined N days and w events, the earthquake sequence θ can be segmented into a set of samples $S = \{S_i | i = 1, 2, \ldots, n\}$ and $S_i = \langle P_i, E_i^m \rangle$. Based on S, the task of earthquake prediction is to predict the magnitude range of the main shock in the future N days-period.

The main challenge for this task is how to extract and combine the explicit and implicit features for accurate earthquake prediction. In addition, the other important challenge is that there usually exists serious category imbalance problem in seismic data. In Section III, we will propose a deep learning model that can effectively deal with these challenges.

B. Convolutional Neural Network (CNN)

Note that the CNN is adopted in our model for implicit feature extracting, therefore, we will introduce the structure of CNN in this part.

Since LeCun et al. [10] proposed the basic architecture of CNN, and successfully applied to handwritten numeral recognition, the CNN began to be widely used in various fields, such as video surveillance, mobile robot vision, image search engine and seismic prediction [11]–[14]. The good performance of CNN lies that convolutional neurons have excellent properties, the convolutional kernel is more efficient than the parametric matrix in the fully connected layer, and does well in extracting structural characteristics of high-dimensional data.

Typical CNN consists of two parts: feature extractor and classifier. The feature extractor is used to filter input data into the "feature map" representing various features and the classifier is used to process the low dimensional vectors from the extractor, and estimate the label of the feature. Fig. 2 shows the calculation process of the convolution layer. In our model, the CNN extractor is made up of multiple computing layers. For example, the extractor may include several convolution layers and optional sub-sampling layers. The convolution layer receives N feature diagram as input, each feature vector convolved through a sliding window with the kernel $K \in R^{k*k}$, to generate a pixel in the output feature diagram. The slide step of the window equals to S, which is often less than k (the size of the convolution kernels), and the mappings of M output features will form the input feature mapping set of the next convolution layer.

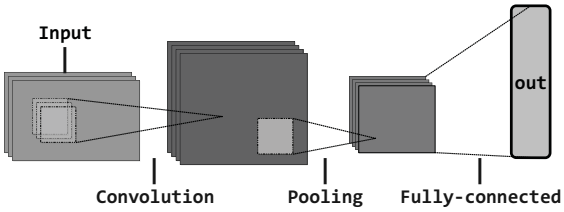

Fig. 2: The structure of CNN

C. Related Work

In the last few decades, many researchers regarded earthquake prediction as a purely geological and physical problem. They tried to discover more effective features and earthquake precursors to predict the future earthquake with the development of physics and geology [1]–[7], [15], [16]. For example, Zhang et al [7] recently proposed a precursory pattern-based feature extraction method for earthquake prediction, where the eight mathematical statistic features can be generated as seismic indicators (i.e. *the time, mean magnitude, seismic root of seismic energy, b-value, mean square deviation, maximum difference, and coefficient of variation*). Compared with different models such as SVM [17], BP [18] and PNN [19], their experimental results on two historical earthquake records demonstrated the effectiveness of their precursory pattern-based features with the selected CART algorithm for earthquake prediction.

Unfortunately, the performance of these methods is usually limited by the characteristics of seismic zones. For example, the work in [6] predicted the earthquake events in Chile with the magnitude larger than 4.4, while the work in [7] only adopted two zones in China. For other seismic data with different properties, previous methods often need some adjustment or even modify the prediction algorithm. To sum up, these seismic indicators (explicit features) designed by humans have strong interpretability from the theoretical system. However, they may fail to fully utilize information contained in seismic sequences. For this purpose, people hope to discover the plentiful features hiding in seismic data.

To this end, some researchers try to establish the neural network model to learn the implicit features directly from the data without explicitly modeling features. The theory of deep learning methods points out another direction for earthquake prediction. The ability of extracting features automatically makes it has been widely applied in academic problems such as image recognition and object recognition. For earthquake prediction, through the training of neural network, the weight vector can describe the characteristics of input data to a certain extent. For example, in work [8], they proposed a deep-learning approach to identify a static-stress-based criterion, with the aim to predict aftershock locations without prior assumptions about fault orientation. Another representative work is [9], where they extracted features by CNN from the map marked with the latitude and longitude of seismic data and used the past 120 days of seismic events to predict the main shock in Taiwan in upcoming 30 days. Note that implicit features extracted by deep learning methods can fully utilize

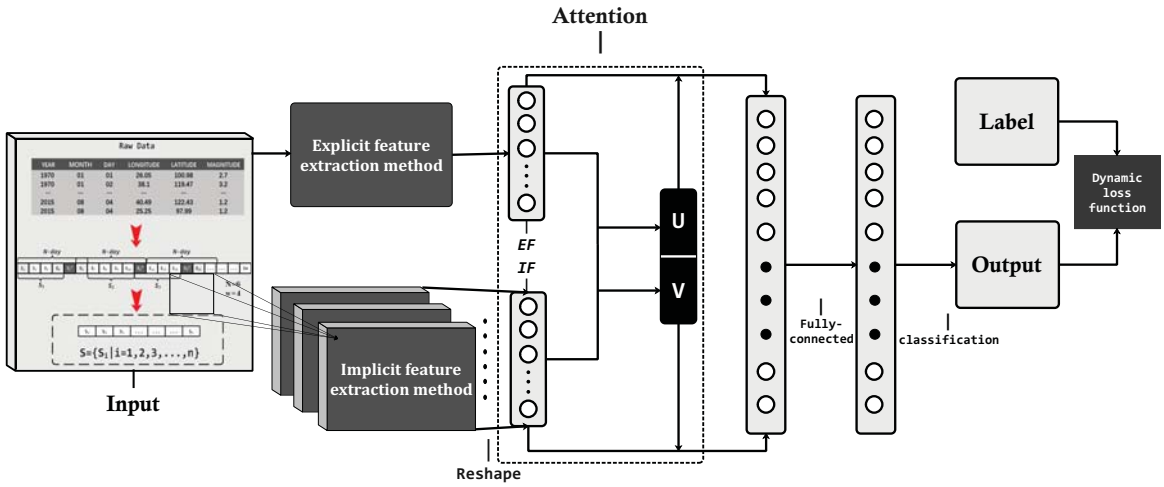

Fig. 3: The framework of DLEP

information contained in seismic sequences, however, they have weak interpretability from the theoretical system.

To sum up, most of the above existing works try to use either seismic indicators (i.e. explicit features) designed by geologists, or use feature vectors (i.e. implicit features) extracted by deep learning methods to characterize the earthquake for earthquake prediction. Different from these works, in this paper, we propose a novel deep learning model by effectively combining the advantages of both explicit and implicit features, thus can greatly improve the performance of earthquake prediction.

III. The Proposed Model DLEP

In this section, we first introduce the general framework of the proposed model DLEP, then give the suggested attention-based strategy for fusing explicit and implicit features, and finally present the proposed dynamic loss function for dealing with the category imbalance problem of seismic data.

A. Overall Framework of DLEP

Fig. 3 gives the general framework of the proposed DLEP, which consists of four steps: data preprocessing, feature extraction, feature fusion and prediction. In the first step of data preprocessing, we use the proposed segment method introduced in Section II-A to extract precursory patterns and training samples. In the second step of feature extraction, we adopt the eight mathematical statistics-based earthquake indicators [7] based on the obtained precursory patterns as the explicit feature vector, denoted as EF. The eight indicators are *the time, mean magnitude, seismic root of seismic energy, b-value, mean square deviation, maximum difference, and coefficient of variation*. In addition, we use CNN to extract implicit vector based on the obtained samples, denoted as IF. In the third step of feature fusion, we suggest an attention-based strategy in Section III-B by using the parameter matrices $\mathbf{U}$ and $\mathbf{V}$ to weight the EF and IF respectively. Then, the fusion vector will be input into the full-connected layer to get the output. During the training phase, the category imbalance problem caused by data distribution tends to cause the model to converge to the local minimum, which is solved by the dynamic loss function proposed in Section III-C. Finally, the model outputs the magnitude range of main shock. More specifically, previous experiments have proved that the ReLU activation function is effective in the CNN, and the *softmax* is often adopted in the fully-connected layer as the activation function, thus we choose them in our model. To enhance the generalization performance of our model, similar to the work in [20], [21], we also adopt dropout layer and batch normalization layer in our model.

From the above explanation of DLEP, it can be found that the proposed attention-based strategy and the dynamic loss function are two important parts in DLEP. In the following, we will illustrate them in detail.

B. Attention Mechanism

In the overall procedure of DLEP, the main challenge is the combination of explicit and implicit features. In this paper, we find that the eight explicit features proposed in [7] and the implicit extracted by CNN have some differences in properties, cannot be simply joined together. Moreover, the CNN has the properties of parameter sharing, which means that it is hard for CNN to focus on the weight of seismic data at different times. But in fact, whether it is an explicit or implicit feature vector, the importance of each dimension is different, such as the indicator of *seismic root of seismic energy* often more decisive than *mean square deviation* in prediction [7]. Therefore, we should not only weight the explicit and implicit features, but also weight each dimension within the vectors. Inspired by the forget gate in long-short term memory(LSTM) [22], we suggest an important attention-based strategy to improve the accuracy and robustness of the prediction task.

To be specific, the attention is used to dynamically weight different features. As shown in Fig. 4, we first simply splice

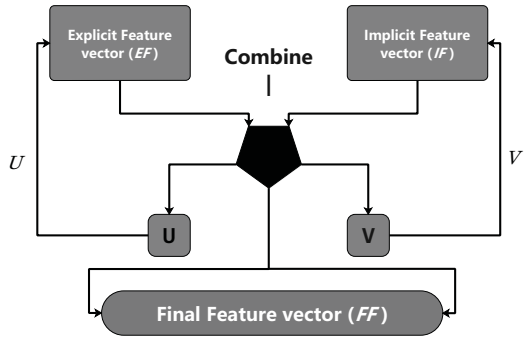

Fig. 4: The structure of Attention Mechanism

the two feature vectors (i.e. explicit feature vector (EF) and implicit feature vector (IF)) and return the fusion vector F_c, whose mathematical expression is given as follows:

$$F_c = [EF, IF]$$

where $EF \in R^{d_1}$ with length d_1 and $IF \in R^{d_2}$ with length d_2 are the feature vectors which extracted from [7] and CNN respectively, $F_c \in R^{d_1+d_2}$ is the merged vector. In the strategy of attention, given the pre-defined parameter matrix $\mathbf{U} \in R^{(d_1+d_2)*d_1}$ and $\mathbf{V} \in R^{(d_1+d_2)*d_2}$, we can get the weight vectors, $U \in R^{d_1}$ and $V \in R^{d_2}$, which defined as:

$$U = sigmoid(F_c \cdot \mathbf{U})$$

$$V = sigmoid(F_c \cdot \mathbf{V})$$

where the $\mathbf{U}$, $\mathbf{V}$ are the two parameter matrices, and the $sigmoid(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, which is used to process the input vector by dimensions. Note that in the weight vector $U \in R^{d1}$ ($V \in R^{d2}$), each dimension represents the weight of the corresponding dimension of the explicit (implicit) feature vector $EF \in R^{d_1}$ ($IF \in R^{d_2}$). Then, the weight vectors U and V will be returned, multiplied by the corresponding EF and IF, and the final fusion feature vector FF is obtained as:

$$FF = [EF \cdot U, IF \cdot V]$$

It can be found that the suggested attention mechanism considers not only the equality of weights caused by parameter sharing problem of CNN in feature extraction, but also the dynamic weights between different features. Thus, the attention-based strategy enhance the generalization ability of our model synchronously, which allows the model to handle seismic data from vastly different seismic zones.

C. Dynamic Loss Function

In the fully-connected layer of DLEP, the category imbalance of raw data will seriously affect model prediction performance. We find that the earthquakes with a magnitude of less than 4.0 account for around 50.44% of all the earthquake events and 86.64% of those with magnitudes less than 5.0 in the seismic data from Sichuan province of China, and the similar situation occurs in other seven seismic datasets adopted in our experiments. Thus, it is difficult to predict large earthquakes. To this end, in this paper, we design a new dynamic loss function in fully-connected layer to accommodate the particularity of seismic data.

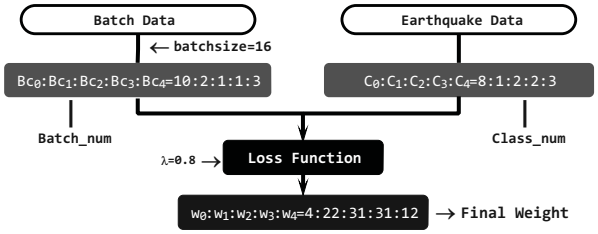

Fig. 5: The structure of Dynamic Loss Function

Specifically, the aim of our dynamic loss function is to weight different samples with different weights according to the distribution characteristics of the training samples. Note that we use the small batch gradient descent method to train the model, therefore, the overall distributions of the sample and the batch should be both considered in loss function. If only the overall distribution of the sample is considered, then there exists difference between the distribution of overall samples and each batch, leading to gradient oscillation of the model. Otherwise, if only the batch distribution is considered, then the model can hard to converge. Therefore, we set a weight parameter λ between the two distributions of the sample and the batch, our dynamic loss function can adaptively modify the parameter value according to the characteristics of different sample set. Formally, our dynamic loss function can be defined as follows:

$$w_i = \lambda * bc_i + (1 - \lambda) * c_i$$

and

$$bc_i = \frac{batch\_size}{Z * batch\_num_i}, Z = \sum_{i=0}^{N} \frac{batch\_size}{batch\_num_i}$$

$$c_i = \frac{data\_num}{Z' * class\_num_i}, Z' = \sum_{i=0}^{N} \frac{data\_num}{class\_num_i}$$

where $batch\_num$ and $class\_num$ are the ratio of the number of samples in each batch and the total sample set respectively, $batch\_size$ and $data\_num$ are the sum of the number of samples in batches and sample set respectively, and Z and Z' are the normalization coefficient respectively. For example, As shown in Fig. 5, suppose $batch\_size$ = 16, $bacth\_num_0$ = 10, $class\_num_0$ = 8 and λ = 0.8, according to the above equation, the weights w_0, w_1, w_2, w_3 and w_4 are 0.0374, 0.2177, 0.3134, 0.3134, 0.1180 respectively. Finally we can get the weight ratio $w_0 : w_1 : w_2 : w_3 : w_4 = 4 : 22 : 31 : 31 : 12$.

Based on this dynamic loss function, our training process considers the sample distribution of each batch and the whole, and the function can effectively guide the small batch gradient descent algorithm to train the model, and finally converges to a better classification model.

TABLE I: The characteristics of eight seismic data sets.

Region	Latitude	Longitude	Instances	Countries	Label-1	Label-2	Label-3	Label-4	Label-5	Instance numbers in each label
Sichuan	28-36N	98-106E	906	China	[3.0,4.0)	[4.0,4.5)	[4.5,5.0)	[5.0,5.5)	[5.5,max)	457 / 214 / 114 / 54 / 67
Xinjiang	35-50N	75-95E	1027	China	[3.0,4.0)	[4.0,4.5)	[4.5,5.0)	[5.0,5.5)	[5.5,max)	123 / 335 / 316 / 142 / 111
Qinghai-Tibet	26-39N	73-104E	1021	China	[3.0,4.0)	[4.0,4.5)	[4.5,5.0)	[5.0,5.5)	[5.5,max)	172 / 208 / 230 / 157 / 254
Shadong-Jiangsu	29-38N	114-124E	658	China	[3.0,4.0)	[4.0,4.5)	[4.5,5.0)	[5.0,5.5)	[5.5,max)	481 / 100 / 49 / 15 / 13
Japan	31-38N	136-143E	1096	Japan	[3.0,4.8)	[4.8,5.3)	[5.3,5.8)	[5.8,6.3)	[6.3,max)	136 / 416 / 317 / 139 / 88
Philippines	11-19N	115-124E	1052	Philippines	[3.0,4.8)	[4.8,5.3)	[5.3,5.8)	[5.8,6.1)	[6.1,max)	67 / 392 / 359 / 112 / 122
Chicago	38-47N	82-93W	610	USA	[3.0,3.2)	[3.2,3.5)	[3.5,4.0)	[4.0,4.5)	[4.5,max)	276 / 99 / 140 / 73 / 22
Los Angeles	30-40N	115-125W	1182	USA	[3.8,4.0)	[4.0,4.5)	[4.5,5.0)	[5.0,5.5)	[5.5,max)	417 / 409 / 196 / 118 / 42

IV. EXPERIMENTS

In this section, we first present experimental settings, including datasets, comparison algorithms and the evaluation metrics. Then, we show the experimental results and analysis.

A. Experimental Settings

1) Data preparation: In this paper, we adopt eight popular seismic zones[1] with different characteristics as our datasets to test the performance of the comparison algorithms. Specifically, the eight zones are Sichuan Province, Xinjiang Province, Qinghai-Tibet plateau, Shandong-Jiangsu Province, Japan, the Philippines, Chicago and Los Angeles. Table I gives the main characteristics of eight seismic data sets, including the number of instances, corresponding longitude and latitude of different regions, belonged countries, the range of earthquake categories and the number of instances in each category. For each dataset, we manually divide the magnitude range into five labels. It is noted that the dividing threshold for each label is slightly different for each dataset, with the aim to get the balanced number of instances for each label. Based on this, we can regard the prediction of earthquake magnitude range as a classification problem.

We use K-fold cross-validation [23] to evaluate the prediction results and k is set to 10 in our experiments, which means that each dataset is divided into 10 parts, nine of which are used as training data, the remaining one as testing data for 10 times.

TABLE II: The main referring parameters used in the DLEP.

Parameters	Value	Annotation
$Epochs$	65	The number of iterations of the entire model
$Batchsize$	24	The number of training samples per batch
$Indim1$	600	The number of neurons in explicit feature layer
$Indim2$	600	The number of neurons in implicit feature layer
lr	0.00025	Learning rate

2) Comparison Algorithms: In order to verify the validity of our model DLEP, we select three state-of-the-art earthquake prediction methods and two variants of DLEP.

- 2016N: The method was proposed in 2016 [24]. It extracted the earthquake indicators before the main shock, and proposed a three-layer feedforward BP neural network to predict earthquakes in the Himalayas.
- 2018CNN: The method was proposed in 2018 [9]. It adopted the model structure of the CNN, and the data before the main shock was removed and marked on the map by latitude and longitude, different magnitudes were marked with different colors and ranges. Finally, the obtained samples were extracted by CNN and used in the earthquake prediction task.
- 2019F: The method was proposed in 2019 [7]. In this method, the original seismic data was segmented according to fixed pre-defined N days, and the precursory pattern was defined based on this. Then, eight features were extracted from the original data, and used the CART for classification.
- DLEP(-Att): In order to show the effectiveness of the proposed attention-based strategy in DLEP, we compare DLEP with one variant of DLEP, namely DLEP(-Att). DLEP(-Att) is the same as DLEP without using the proposed attention-based strategy.
- DLEP(-Dyn): In order to show the effectiveness of the proposed dynamic loss function in DLEP, we compare DLEP with one variant of DLEP, namely GTEA(-Dyn). GTEA(-Dyn) is the same as DLEP without using the proposed dynamic loss function.

For fair comparisons, we adopt the recommended parameters values for all the comparison algorithms, which were suggested in their original papers. For the proposed DLEP and two variants DLEP(-Att) and DLEP(-Dyn), the main referring parameter settings are given in the Table II. It is worth mentioning that these parameters are just referring values, the real parameters for different dataset may be slightly different. Specifically, in our experimental process, we adopt Bayesian search [25] to conduct hyperparametric search in each experiment, ensuring that each model could achieve the optimal performance of their existing structures.

3) Evaluation Metrics: In this paper, two well-known metrics ACC and $MAUC$ are used to evaluate the performance of comparison algorithms.

ACC is commonly used to show the classification performance of various classifiers, which is defined as follows:

$$ACC = (TP + TN)/(TP + TN + FP + FN)$$

where TP is means true positive, FP means false positive, TN means true negative and FN means false negative.

$MAUC$ is a popular evaluation indicator for evaluating multi-classification problems [26], which is defined as follows:

$$MAUC = \frac{2}{c * (c-1)} \sum_{i<j} \frac{A_{ij} + Aji}{2}$$

where the A_{ij} is the AUC value [27] between class i and class j, which is calculated based on the i-th and j-th column

[1]https://earthquake.usgs.gov/earthquakes/search/

TABLE III: The experimental results of comparison methods in terms of ACC and $MAUC$ on eight datasets.

Region	ACC				$MAUC$			
	2016N	**2018CNN**	**2019F**	**DLEP**	**2016N**	**2018CNN**	**2019F**	**DLEP**
Sichuan	57.71	38.41	52.32	**81.23**	61.00	51.23	71.81	**90.07**
Xinjiang	32.71	34.38	68.76	**77.52**	55.39	50.87	74.19	**92.42**
Qinghai-Tibet	34.15	31.85	62.47	**75.69**	57.33	49.98	68.19	**91.41**
Shadong-Jiangsu	67.08	82.03	54.85	**84.13**	**86.64**	50.00	67.64	71.55
Japan	52.68	28.33	68.16	**79.42**	69.20	50.10	73.59	**90.23**
Philippines	74.37	32.45	71.41	**77.83**	66.42	50.41	63.52	**89.46**
Chicago	47.42	40.07	76.08	**76.36**	56.84	51.04	60.23	**85.02**
Los Angeles	35.26	35.48	56.36	**79.88**	58.68	50.67	72.67	**89.81**

in $M \in R^{n*c}$, and n and c are the number of instances and classes respectively.

For both ACC and $MAUC$, the larger value indicates the better performance for classification.

B. Experimental Results

TABLE IV: The experimental results of DLEP and its two variants DLEP(-Att) and DLEP(-Dyn) in terms of $MAUC$ on eight datasets.

Region	$MAUC$		
	DLEP	**DLEP(-Att)**	**DLEP(-Dyn)**
Sichuan	**90.07**	77.04	55.58
Xinjiang	**92.42**	78.86	60.84
Qinghai-Tibet	**91.41**	79.21	88.07
Shadong-Jiangsu	**71.55**	60.29	52.79
Japan	**90.23**	79.65	76.87
Philippines	**89.46**	80.63	78.54
Chicago	**85.02**	77.19	51.27
Los Angeles	**89.81**	80.78	75.43

1) Effectiveness of The Proposed DLEP: Table III presents the experimental results of comparison methods in terms of ACC and $MAUC$ on eight datasets, where the best performance for each dataset is marked with bold. From this table, it can be found that the proposed DLEP gets the best ACC among comparison algorithms on all datasets. The good performance of DLEP is attributed to the fact that DLEP can effectively combine the advantages of both explicit features and implicit features for earthquake prediction, while 2016N and 2019F are only use explicit features, and 2018CNN only use implicit features. We can also find that baseline methods tend to work well with certain datasets, but badly with others. For example, the ACC of baseline 2018CNN on Shandong-Jiangsu dataset works well and reaches 82.03%, however, for dataset Japan, it works badly and just reaches 28.33%. This is mainly due to the huge geological differences in the zones, which indicates that the generalization ability of DLEP is better than that of other baseline models.

Similarly, it can also be found in Table III that the proposed DLEP obtains the best $MAUC$ among comparison algorithms on seven datasets, and get the second best on the remaining one dataset (Shandong-Jiangsu dataset). Let us take a close look at Shandong-Jiangsu dataset, the $MAUC$ of 2016N reaches the best 86.64% while that of DLEP reaches the second best 71.55%. But compared to the ACC value under the same conditions, 2016N just reaches 67.08% while DLEP reaches 84.13%. The reason behind this is that 2016N can predict earthquakes with lower magnitude well on Shandong-Jiangsu dataset. Therefore, the effectiveness of the proposed model MOEA-DIM compared with baselines is verified.

2) Effectiveness of The Proposed Strategies used in DLEP: In this section, we will verify the effectiveness of the proposed strategies used in MOEA-DIM, that is, the attention-based strategy for fusing both explicit and implicit features and a dynamic loss function for dealing with the category imbalance of seismic data. Table IV shows the experimental results of DLEP and its two variants DLEP(-Att) and DLEP(-Dyn) in terms of $MAUC$ on eight datasets. Note that the dynamic loss function is designed to improve the $MAUC$, it is meaningless to show ACC of DLEP without dynamic loss function. In addition, the metric of $MAUC$ is relative better than ACC used in our problem, which is a multi-classification problem. Thus, we give the comparison results focused on metric $MAUC$.

It can be observed from this table that the proposed DLEP is greatly better than DLEP(-Att), about 14% improvement. In order to better show that advantages of the proposed attention-based strategy, we adopt t-SNE [28] to visualize the feature layer after attention in the overall procedure of DLEP, and the results are shown in Fig. 6. Due to space limitation, we only show the results on two datasets, Sichuan and Qinghai-Tibet, the similar results can also be found on other datasets. From this figure, we can find that DLEP with the strategy of attention can increase the distance between samples of different labels in the sample space, making them easier to be distinguished. In addition, we can find that the proposed DLEP is greatly better than DLEP(-Dyn), about 29% improvement. It can be concluded that DLEP with both the proposed attention-based strategy and the dynamic loss function can greatly improve the final prediction performance, therefore the effectiveness of the proposed two strategies are verified.

V. Conclusion and Future Work

In this paper, we argued that the feature extraction methods used in previous work obtain either explicit features designed by geologists, or implicit features extracted by deep learning methods, and lack a general model that can fuse the advantages of both explicit and implicit features. To this end, a deep learning model named DLEP is proposed to combine the explicit and implicit features for accurate earthquake prediction. In our model, eight precursory pattern-based indicators were adopted as the explicit features, while a convolutional neural network (CNN) was adopted to extract implicit features. After that, an attention-based strategy was suggested to fuse these

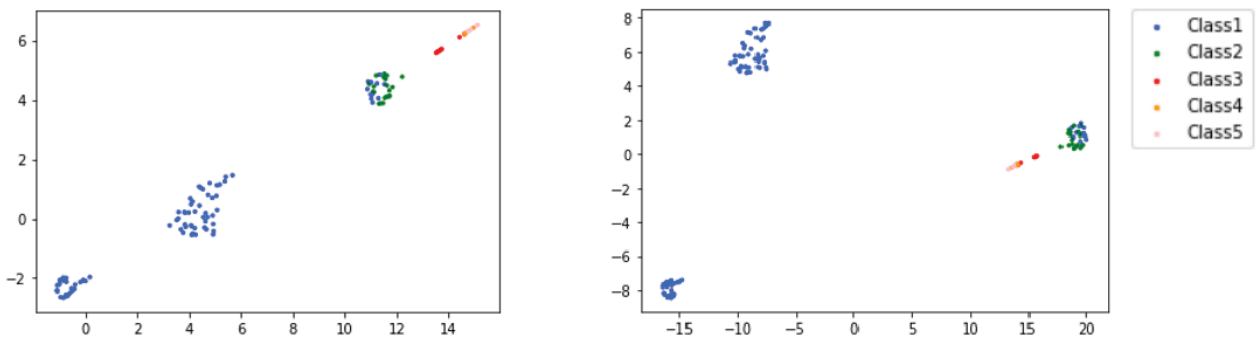

Fig. 6: The visualization results on Sichuan (left part) and Qinghai-Tibet (right part) datasets. Noting that the practical significance of horizontal and vertical axis cannot be explained, this figure only represent the distribution of samples in the sample space.

two kinds of features well. Moreover, a dynamic loss function was also designed to solve the category imbalance problem of seismic data. Finally, compared with several state-of-the-art baselines, the experimental results on eight datasets with different characteristics demonstrated the promising performance of the proposed DLEP for accurate earthquake prediction. It is worth noting that in the proposed DLEP, the explicit features and implicit features are only adopted from existing work, in the future, we would like to design or extract more effective explicit features and implicit features to further improve the earthquake prediction performance.

ACKNOWLEDGEMENT

This work is supported by the Natural Science Foundation of China (Grant No.61976001 and 61876184), the Natural Science Foundation of Anhui Province (1908085MF219), and the Humanities and Social Sciences Project of Chinese Ministry of Education (Grant No.18YJC870004).

REFERENCES

[1] C. H. Scholz, L. R. Sykes, and Y. P. Aggarwal, "Earthquake prediction: a physical basis," *Science*, vol. 181, no. 4102, pp. 803–810, 1973.

[2] H. Kanamori and D. L. Anderson, "Theoretical basis of some empirical relations in seismology," *Bulletin of the seismological society of America*, vol. 65, no. 5, pp. 1073–1095, 1975.

[3] S. Crampin, R. Evans, and B. K. Atkinson, "Earthquake prediction: a new physical basis," *Geophysical Journal International*, vol. 76, no. 1, pp. 147–156, 1984.

[4] B. Gutenberg and C. F. Richter, "Earthquake magnitude, intensity, energy, and acceleration," *Bulletin of the Seismological society of America*, vol. 32, no. 3, pp. 163–191, 1942.

[5] C.-Y. King, "Gas geochemistry applied to earthquake prediction: An overview," *Journal of Geophysical Research: Solid Earth*, vol. 91, no. B12, pp. 12 269–12 281, 1986.

[6] E. Florido, F. Martínez-Álvarez, A. Morales-Esteban, J. Reyes, and J. L. Aznarte-Mellado, "Detecting precursory patterns to enhance earthquake prediction in chile," *Computers & geosciences*, vol. 76, pp. 112–120, 2015.

[7] L. Zhang, L. Si, H. Yang, Y. Hu, and J. Qiu, "Precursory pattern based feature extraction techniques for earthquake prediction," *IEEE Access*, vol. 7, pp. 30 991–31 001, 2019.

[8] P. M. DeVries, F. Viégas, M. Wattenberg, and B. J. Meade, "Deep learning of aftershock patterns following large earthquakes," *Nature*, vol. 560, no. 7720, pp. 632–634, 2018.

[9] J. Huang, X. Wang, Y. Zhao, C. Xin, and H. Xiang, "Large earthquake magnitude prediction in taiwan based on deep learning neural network," *Neural Network World*, vol. 28, no. 2, pp. 149–160, 2018.

[10] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Efficient deep cnn-based fire detection and localization in video surveillance applications," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 7, pp. 1419–1434, 2018.

[12] L. Carnimeo, "A cnn-based vision system for pattern recognition in mobile robots," in *Proc. of the 15th IEEE European Conf. on Circuit Theory & Design, Espoo, Finland*, 2001.

[13] Y. Li, H. Su, C. R. Qi, N. Fish, D. Cohen-Or, and L. J. Guibas, "Joint embeddings of shapes and images via cnn image purification," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, p. 234, 2015.

[14] Y. Yu, J. Lin, L. Zhang, G. Liu, J. Hu, Y. Tan, and H. Zhang, "Identification of seismic wave first arrivals from earthquake records via deep learning," in *Proceedings of 11th International Conference on Knowledge Science, Engineering and Management*, 2018, pp. 274–282.

[15] H. Adeli and A. Panakkat, "A probabilistic neural network for earthquake magnitude prediction," *Neural networks*, vol. 22, no. 7, pp. 1018–1024, 2009.

[16] A. Panakkat and H. Adeli, "Neural network models for earthquake magnitude prediction using multiple seismicity indicators," *International journal of neural systems*, vol. 17, no. 01, pp. 13–33, 2007.

[17] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Oxford university press, 1995.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[19] D. F. Specht, "Probabilistic neural networks," *Neural networks*, vol. 3, no. 1, pp. 109–118, 1990.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[22] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[23] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of machine learning research*, vol. 5, no. Sep, pp. 1089–1105, 2004.

[24] S. Narayanakumar and K. Raja, "A bp artificial neural network model for earthquake magnitude prediction in himalayas, india," *Circuits Syst*, vol. 7, no. 11, pp. 3456–3468, 2016.

[25] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[26] K. Tang, R. Wang, and T. Chen, "Towards maximizing the area under the roc curve for multi-class classification problems," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[27] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Machine learning*, vol. 45, no. 2, pp. 171–186, 2001.

[28] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.